Linked Data and Spatial Data Infrastructures

Simon Cox[1,2]
[1]CSIRO, Kensington, WA, Australia, simon.cox@csiro.au
[2]European Commission Joint Research Centre, Ispra (Va), Italy, simon.cox@jrc.ec.europa.eu

INTRODUCTION

Spatial Data Infrastructures (SDIs), such as AuScope Grid [1] and the EU's INSPIRE [6, 7], are being planned and built using discovery, access and processing components based on a services model. While the principle of distribution and delegation using the internet is a major step forward from traditional data warehouses and private collections, the query-oriented interaction paradigm is merely evolutionary compared with traditional access systems designed for expert users. In contrast, the success and scalability of the world wide web has been based on hypertext, in which browsing is the key mode of interaction, supported by Universal Resource Identifiers (URIs).

Linked Data [2] has been proposed as the bridge from the browseable web to the deep web of technical data, organizing it as a graph. Linked Data is still based on web-pages (usually HTML) for user interactions, but supported by Resource Description Framework (RDF) for richer link semantics. Importantly, links are expected to frequently resolve to data in legacy file formats which serve as leaf-nodes, but in this way are made part of the web of resources. This has led to Linked Data being proposed in some influential jurisdictions as a panacea for publication of information managed by statutory agencies which is currently hard to access, including geospatial data [3, 9].

However, key standards used in SDI were designed on Linked Data principles, even before the name was coined. For instance, Geography Markup Language (GML) is essentially an RDF application, allowing for links embedded within data, with semantics provided by the name of the link element [11, 12]. In principle, SDIs should integrate seamlessly into the web of linked data. There are, however, a number of issues to consider or resolve in order to bring this about.

WHAT ARE THE RESOURCES?

The basic assumption of the Linked Data paradigm is linking to documents and to complete data-sets. While this may be appropriate, and often the only possibility, for file-oriented data, most geospatial data is organized in databases. A subset useful for a particular study is extracted by a parameterized query. The SDI service interface standards (WMS, WFS, SOS, WCS [10]) provide for the query to be encoded in a URL. Each new query implicitly creates a new 'resource', so a large, possibly infinite, set of distinct resources is associated with each database.

Because of this the interaction resource must provide a UI for query construction. But in this scenario, the links, and thus resources, are associated with a web application which is probably decoupled from the data-service itself. This is distinct from the conventional 'browse' metaphor, based on a list of links which refer to persistent resources defined and hosted in association with the data custodian. In this context, a directly usable representation of the service 'capabilities' statement in the form of a navigation resource would allow the standard SDI interfaces to fit better into the Linked Data world (and the mass-market web in general).

REPRESENTATIONS OR RESOURCES?

The Linked Data approach is based in REST principles and techniques [2, 8], in which content negotiation may be used for alternative resource representations. In the http protocol, selection of representation is on the basis of the MIME-type. However, under the domain-modeling principles used in the SDI community, the schema for reporting geospatial features is application-specific. For example, a road may be modeled in different communities as a link in a transportation network, as an engineering structure, or as an ecosystem boundary, etc. So the same real-world feature may be available from different (domain-specific) servers with different sets of properties, though all encoded using GML.

The identifiers disambiguate these representations. They act as different information-resources, but all associated with a single real-world (non-information) resource. Linked Data is adept at handling the case of multiple resources associated with a master URI (for the real-world resource), but this requires authoritative maintenance of persistent identifiers, perhaps as part of a gazetteer service.

STRUCTURED DATA AND LINKED DATA

Linked Data has focused on web-pages (primarily HTML) for human browsing, and RDF for semantics, assuming that other representations are opaque. However, this overlooks the wealth of XML data on the web, which has supported the mash-up phenomenon. Some is structured according to XML Schemas that provide semantics.

Technical applications can use content-negotiation to get a structured representation, so as to exploit the semantics. For example, in the earth and environmental sciences there is increasing use of data representations based on GML, such as GeoSciML, O&M, WaterML, AIXM and MOLES. GML was strongly influenced by RDF, and typed links are intrinsic: xlink:href plays the role that rdf:resource does in RDF representations, so GML data can also serve as a Linked Data navigation resource. Furthermore, many GML application schemas are derived directly from information models formalized using a UML profile that is easily transformable into RDF [3] and thus provides fairly rigorous semantics, though with some expressivity limitations, so semantics may be 'lifted' from the data. Services which expose GML-formatted resources (such as OGC Web Feature Service, WFS) are a prototype of Linked Data.

VOCABULARIES AND VOCABULARY SERVICES

Within structured datasets, semantic interoperability is achieved most easily by them using common sources for the content as well as structure. For example, within the geological sciences there is a long history of controlled vocabulary development and publication, supported by the stable institutional arrangements embodied in the Geological Surveys. Transforming these into semantic-web-friendly form (SKOS, RDF) is relatively straightforward [5, 13]. However, one of the key strengths of these resources is their orderly maintenance regime. Unfortunately this is an area where there has been insufficient attention in RDF technology, and standards for the recording of provenance and status have lagged in the current generation of triple-stores. The 'open world' assumption of the semantic web has led to the development of practices that are in some tension with the desire of some communities to agree to use 'standard' resources from recognized authorities.

CONCLUSIONS

Linked Data principles are compatible with SDI technologies that are currently under deployment in a number of jurisdictions. Some spatial data is naturally fitted to the identifier-orientation of Linked Data, particularly feature services that use GML for data transport. Lifting semantics from GML data should be straightforward. Gridded-data (maps, geophysics, numerical models) that is usually accessed as subsets generated by querying is less well suited for direct linking, but a linked data interpretation can be mediated by web-client interfaces. Services for vocabularies and identifiers for real-world features that are a pre-requisite for SDIs are highly amenable to deployment using semantic-web technologies.

ACKNOWLEDGEMENTS

My interest in Linked Data was developed particularly through conversations with Sven Schade, Andrew Woolf, Clemens Portele, Keith Murray and Stuart Williams.

REFERENCES

1. *AuScope Grid and Interoperability* http://auscope.org/content.php/content/id/12 accessed 27 June 2010
2. Bizer, C., T. Heath, and T. Berners-Lee, *Linked Data: Principles and State of the Art.* World Wide Web Conference, 2008. www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf
3. Compton, M., Neuhaus H., Bermudez L., Cox S.J.D., *An Ontology for Sensor Networks*. Geophysical Research Abstracts **12**, Proceedings EGU General Assembly 2010 http://meetingorganizer.copernicus.org/EGU2010/EGU2010-3817-1.pdf
4. DEFRA (UK), *UK Location Program*. http://location.defra.gov.uk/
5. Githaiga, J., Duclaux, G., Cox, S.J.D., & Yu, J. *Spatial Information Services Stack (SISS) Vocabulary Service – A Tool For Managing Earth & Environmental Sciences Controlled Vocabularies.* This conference.
6. INSPIRE, *INSPIRE scoping paper. http://www.ec-gis.org/inspire/reports/inspire_scoping24mar04.pdf, 2004.*
7. *Infrastructure for Spatial Information in the European Community (INSPIRE)* http://inspire.jrc.ec.europa.eu/ accessed 27 June 2010
8. Jacobs, I. & Walsh, N. (eds). *Architecture of the World Wide Web, Volume One (2004)* http://www.w3.org/TR/webarch/
9. Murray, K., et al. *Linked Data and INSPIRE – extending the benefits of data sharing. INSPIRE Conference 2010 http://inspire.jrc.ec.europa.eu/events/conferences/inspire_2010/conf_skd_conference.cfm*
10. Percivall, G. et al, OGC Reference Model (ORM) - Version 2.0. The Open Geospatial Consortium, 2008. http://www.opengeospatial.org/standards/orm
11. Portele, C. (ed) *OpenGIS Geography Markup Language (GML) Encoding Standard - Version 3.2.1*. The Open Geospatial Consortium, 2007. http://www.opengeospatial.org/standards/gml
12. Schade, S. & Cox, S.J.D. *Linked Data in SDI or How GML is not about Trees*. AGILE 2010
13. Yu J., Cox S.J.D., Ratcliffe, D., *Use of Standard Vocabulary Services in Validation of Water Resources Data*. Geophysical Research Abstracts **12**, Proceedings EGU General Assembly 2010 http://meetingorganizer.copernicus.org/EGU2010/EGU2010-7577.pdf

ABSTRACT

Spatial Data Infrastructures (SDIs) such as AuScope Grid and INSPIRE are being planned and built using discovery, access and processing components based on a services model. While the principle of distribution and delegation using the internet is a major step forward from traditional data warehouses and private collections, the query-oriented interaction paradigm is merely evolutionary, compared with traditional access systems designed for expert users.

In contrast, the success and scalability of the world wide web has been based on hypertext, in which browsing is the key mode of interaction, supported by Universal Resource Identifiers (URIs). Linked Data has been proposed as the bridge from the browseable web to the deep web of technical data. Linked Data is still based on web-pages (usually HTML) for user interactions, but supported by Resource Description Framework (RDF) for richer link semantics. Links can resolve to datasets in legacy file formats which thus serve as leaf-nodes, but can also be part of the web of resources.

Key standards used in SDI were designed on Linked Data principles, even before the name was coined. For instance, Geography Markup Language (GML) is essentially an RDF/XML application. Thus, in principle, SDIs should integrate seamlessly into the web of linked data. There are, however, a number of issues to consider or resolve in order to bring this about. These include: the definition of 'resource' in the context of databases that are accessed as projected subsets by query, accessed through web-service interfaces; multiple representations of the same feature from different services to support different applications; semantics embedded in structured representations expressed in non-RDF XML forms; standard vocabulary and identifier services.

BIO

Simon Cox trained in geology and geophysics, starting his research career on laboratory studies of the mechanical properties of rocks. He developed an interest in visualization and information management, with early work in web-mapping and metadata systems, before focusing on geospatial information standards. He is temporarily based at the Joint Research Centre of the European Commission in northern Italy, working on harmonisation of international geospatial data initiatives. At CSIRO he most recently was project leader for Water Data Transfer Standards, and standards liaison for AuScope GRID (NCRIS).

Dr Cox is a member of AGU, EGU, a councillor of IAMG, and previously served on the council of the IUGS Commission for Geoscience Information, and the Dublin Core Advisory Council.

He is editor of *ISO 19156* and chair of the OGC Naming Authority.

Dr Cox was awarded the 2006 Gardels Medal by the Open Geospatial Consortium.